Estadísticas para repositorios: sistema métrico de datos en *Digital.CSIC*

Por **Isabel Bernal** y **Julio Pemau-Alonso**

Resumen: *El acceso abierto a la información ha traído consigo la multiplicación de datos susceptibles de ser aprovechados para análisis estadísticos del impacto de las publicaciones científicas, abriendo la puerta a nuevos modelos métricos de la comunicación científica. En mayo de 2010 el repositorio institucional Digital.CSIC inauguró un nuevo módulo de estadísticas que responden a la complejidad organizativa del CSIC y a las peticiones por parte de sus bibliotecas de generar informes estadísticos por centros. Con estos informes más elaborados se podrán llevar a cabo funciones de seguimiento interno y de promoción y divulgación externa que reflejen con mayor claridad la relación coste-beneficio del repositorio, así como el impacto de la producción científica del CSIC disponible en abierto desde el mismo. Se explica la arquitectura de las estadísticas a la carta elaboradas por Digital.CSIC, así como las necesidades y exigencias que pretenden cubrir.*

Palabras clave: *Digital.CSIC, Repositorios, Estadísticas, Granularidad, DSpace, Métrica, Impacto, Comunicación científica.*

Title: Statistics for repositories: a metric system for data on *Digital.CSIC*

Abstract: *Open access to information has brought about a multiplication of data that can be used to carry out statistical analysis on the impact of scientific publications, thus paving the way for new metrics models for scholarly communication. In May 2010 institutional repository Digital.CSIC launched a new statistics model to address CSIC structural complexity as well as requests by CSIC libraries to generate statistical reports by centres. Thanks to these newly developed statistics, it will be possible to undertake both internal follow-up activities and external promotional and advocacy efforts which will show cost-benefit relationship at Digital.CSIC more clearly alongside the impact of CSIC science that is openly available through Digital.CSIC. The article explains the architecture behind statistics a la carte developed by Digital.CSIC and needs and challenges that are to be addressed.*

Keywords: *Digital.CSIC, Repositories, Statistics, Granularity, DSpace, Metrics, Impact, Scientific communication.*

Bernal, Isabel; **Pemau-Alonso, Julio**. "Estadísticas para repositorios: sistema métrico de datos en *Digital.CSIC*". *El profesional de la información*, 2010, septiembre-octubre, v. 19, n. 5, pp. 534-543.

DOI: 10.3145/epi.2010.sep.15

Isabel Bernal *coordina desde enero de 2010 el repositorio institucional Digital.CSIC. Ha trabajado en la cooperación internacional y la promoción de recursos electrónicos y aplicaciones tecnológicas en bibliotecas, primero en la DG de la Sociedad de la Información de la Comisión Europea y posteriormente durante 5 años en la fundación eIFL.net (Electronic Information for Libraries), donde realizó proyectos con consorcios de bibliotecas en 47 países en desarrollo y en transición. Tiene un máster en biblioteconomía y documentación en la Escuela de la Biblioteca Vaticana y un máster en economía y relaciones internacionales en la Johns Hopkins University.*

Julio Pemau-Alonso *es desde 2009 analista de sistemas de la Unidad de Coordinación de Bibliotecas del Consejo Superior de Investigaciones Científicas (CSIC) y responsable técnico del repositorio institucional Digital.CSIC. Realiza tareas de programador, consultor y analista de diversos proyectos enfocados al mundo Web en la Secretaría General Adjunta de Informática del CSIC. Anteriormente participó como programador y director técnico en distintos proyectos webs en la empresa privada.*

1. Complejidad estructural y riqueza productiva del *CSIC*

***DIGITAL.CSIC* ES UN REPOSITORIO MULTIDISCIPLINAR creado en enero de 2008 para organizar, difundir y preservar los resultados de la investigación realizada en cada uno de los institutos del *Consejo Superior de Investigaciones Científicas (CSIC)*.**

Desde la firma de la *Declaración de Berlín* en 2006 el *CSIC* está comprometido con el movimiento del acceso abierto y *Digital.CSIC* es la materialización institucional de la llamada vía verde. Nace bajo los auspicios de la *Unidad de Coordinación de Bibliotecas* y es una iniciativa innovadora que fomenta la colaboración activa de bibliotecarios e investigadores de la institución, comprometiendo la colaboración de toda su *Red de Bibliotecas* en la actividad de difundir la ciencia producida en el *CSIC*.

Uno de los grandes incentivos en la creación del repositorio fue la ingente cantidad de trabajos de in-

vestigación científica que el *CSIC* ha ido produciendo desde su creación en 1939. Con casi 9.600 personas[1] (incluyendo científicos de plantilla, investigadores contratados y becarios) dedicadas a actividades relacionadas con la investigación, distribuidas en 128 centros de investigación (77 propios y 51 mixtos) repartidos por toda la geografía española y en el extranjero, el *CSIC* es la agencia estatal científica líder en España. Su compleja estructura es el resultado del crecimiento experimentado a lo largo de los años como institución pública, y refleja la evolución de la investigación científica nacional.

http://digital.csic.es/

> *"Digital.CSIC es un repositorio que organiza, difunde y preserva los resultados de la investigación realizada en los institutos del CSIC"*

Entre 2002 y 2007 el *CSIC* generó más de 60.000 publicaciones científicas, excluyendo el material de divulgación. El *Plan de actuación CSIC 2010-2013* prevé que en los próximos 3 años sus centros e institutos publicarán más de 36.000 artículos[1].

A la capacidad investigadora del *CSIC* se une el reconocimiento internacional por su calidad científica; así, aparece entre los primeros puestos en dos conocidas clasificaciones de centros de investigación: en el *Scimago Institutions Rankings 2009*, el CSIC es el 1 centro de investigación más valorado entre los primeros 2.000 del mundo mientras que el *Ranking Webometrics 2010* lo coloca en el puesto 19 sobre los 4.000 primeros centros de investigación mundiales.

http://www.scimagoir.com/

http://research.webometrics.info/top4000_r&d_es.asp

Sólo en 2009 el *CSIC* generó 26.992 publicaciones[2]

9.754	artículos *SCI-SSCI-AHCI*		
1.962	artículos no *SCI-SSCI-AHCI*		
368	libros		
1.784	capítulos de libros		
104	monografías		
4.634	comunicaciones y	3.409	posters en congresos internacionales
2.384	comunicaciones y	1.618	posters en congresos nacionales
795	tesis doctorales y	180	patentes

Digital.CSIC en cifras [enero 2008 - septiembre 2010][3]

Contiene 25.000 registros que reflejan el carácter multidisciplinar de la investigación. La progresión en el crecimiento de contenidos es constante, y el año 2009 se cerró con un incremento del 42% con respecto a 2008.

Contiene artículos post-print (referenciados por pares) y pre-print (aún no evaluados por comisiones de pares), comunicaciones de congresos, informes técnicos, memorias, documentos de trabajo, bases de datos, libros y capítulos de libros, presentaciones, material de divulgación, material audiovisual, mapas, partituras, etc.

3.340.000 visitas a la web de Digital.CSIC.

3.350.000 descargas de textos completos, con los usuarios de los Estados Unidos a la cabeza.

2. Estadísticas de repositorios: una aproximación

2.1. Nuevas perspectivas en la recolección de datos

El acceso abierto a la información ha traído consigo la multiplicación de datos susceptibles de ser aprovechados para análisis estadísticos sobre las publicaciones científicas, abriendo así la puerta a

nuevos modelos métricos que pretenden ser más englobadores –agregando datos procedentes de fuentes de información diferentes– y más granulares –permitiendo llegar a un nivel de detalle en los análisis estadísticos hasta ahora desconocido (**Aguillo**, 2009)–.

Los datos objeto de análisis estadísticos pueden referirse al impacto de la investigación desde el punto de vista de los autores (citaciones y factor de impacto de las revistas académicas) o al uso por parte de los internautas (visitas y descargas de publicaciones). El primer sistema está ampliamente consolidado, tal y como demuestra la difusión del *Impact Factor* (o *Journal Impact Factor*, *JIF*) como criterio usado por las agencias financiadoras y los propios investigadores para evaluar la calidad de la producción científica. Sin embargo, su premisa –que un alto índice de citaciones de una revista equivale a un alto impacto/calidad de un artículo– ha sido cada vez más criticada ya que excluye a muchas publicaciones científicas, da origen a una correspondencia mecánica entre la calidad de una revista y la de los trabajos y autores que en ella publican. Además, este parámetro de calidad no dice mucho a los usuarios. Como se sabe, han surgido índices alternativos como el *Hirsch index* (*h-index*), que centran su evaluación en la producción del autor y no en el índice de impacto de las revistas.

El otro tipo de evaluación alternativa lo aporta el análisis estadístico del uso (visitas y descargas), centrado en el usuario y aplicado sobre todas sus publicaciones, creando así un modelo más general (**Herb**; **Kranz**; **Leidinger**; **Mittlesdorf**, 2010).

Los datos necesarios para el análisis pueden extraerse de los logs de acceso, los resolvedores de enlaces, los motores de búsquedas o directamente, mediante identificadores persistentes de las plataformas donde se encuentra la información objeto de estudio. Pueden también ser recogidos manual o automáticamente. La primera complicación se encuentra en la cantidad y los formatos de datos (en html, xml, cvs, pdf...), que pueden variar dependiendo de la modalidad de recolección escogida.

En relación con las estadísticos de uso, el incremento en la variedad y cantidad de datos derivado de las posibilidades que ofrecen las plataformas en abierto ha generado un intenso debate en los últimos años, por un lado, para llegar a un consenso sobre qué datos recolectar, y por otro para definir cómo realizar los estudios y cómo armonizarlos con otros sistemas de recolección de datos (principalmente los de los editores tradicionales), con el fin de obtener análisis agregados.

"Se puede analizar el impacto de una investigación por las citas de otros autores o por las veces que se descarga"

El debate ha dado lugar a diversos esfuerzos para establecer nuevos estándares y protocolos de alcance internacional que faciliten una verdadera interoperabilidad, para lo que se necesita una uniformización de descriptores de objetos además de pautas comunes en la captura y transferencia de datos. Los estándares *OAI-PMH* (*Open archives initiative protocol for metadata harvesting*), así como los de *OAI-ORE* (*Open archives initiative object reuse and exchange*), *OpenURL context object*, *DOI* (*Digital object identifier*), *Counter* (*Counting online usage of networked electronic resources*) y *Sushi* (*Standardized usage statistics harvesting initiative*, usado para el intercambio de datos en la comunidad de editores) se consideran puntos de partida sobre los que realizar mejoras para la elaboración de un posible estándar internacional (**Scholze**, 2007).

http://www.openarchives.org

http://www.openarchives.org/ore

http://www.niso.org/kst/reports/standards?step=2&project_key=d5320409c5160be4697dc046613f71b9a773cd9e

http://www.doi.org/

http://www.projectcounter.org/

http://www.niso.org/workrooms/sushi

Los datos recogidos en plataformas de acceso abierto ofrecen mayores posibilidades de granularidad o detalle en los resultados estadísticos. La posibilidad de obtener estadísticas de uso basadas en artículos (y en items individuales en general) y no simplemente en títulos de publicaciones se considera un avance importante para analizar el impacto y el uso de la producción científica mediante estudios complejos que giren en torno a autores (impacto, calidad, popularidad, tendencias en líneas de investigación y en pautas de acceso y búsqueda) y usuarios (pautas de acceso y de búsqueda, intereses temáticos, usos, etc.).

"*OAI-PMH*, *OAI-ORE*, *OpenURL context object*, *DOI*, *Counter* y *Sushi* son puntos de partida para la elaboración de un estándar internacional"

En los últimos años diversas iniciativas han explorado posibles estándares de alcance internacional para facilitar la creación e intercambio de estadísticas más complejas, y la integración y la interoperabilidad con los datos estadísticos generados por los repositorios. El

proyecto *Pirus (Publisher and institutional repository statistics)* del *JISC*, finalizado en 2009, demostró que técnicamente es posible crear, registrar y consolidar estadísticas de uso para artículos individuales usando datos de repositorios y editores, abogando para ello por el enriquecimiento de las estadísticas de *Counter* (**Shepherd**, 2009). El proyecto recomendó *OpenURL context object* como el componente esencial para que los repositorios pudiesen describir y transmitir sus estadísticas de uso (**Shepherd**; **Needham**, 2009).

Varias plataformas como *PLoS* y *SURF* ya están aplicando el esquema xml propuesto por *Pirus* para elaborar estadísticas a nivel de artículos. La secuela del proyecto, *Pirus2*, promueve la estrecha colaboración entre *Counter*, *CrossRef*, editores, repositorios, *NISO* e iniciativas similares en Europa y otras zonas para establecer una infraestructura y llevar a cabo una serie de programas en acceso abierto que promuevan la generación y el intercambio de estadísticas de uso de tipo *Counter*. Se trata que engloben a artículos depositados en repositorios y basadas en los items de los repositorios.

Mesur (Metrics from scholarly usage of resources) es un sofisticado proyecto de *Los Alamos National Laboratory* que analiza el impacto de la producción científica proponiendo datos relacionales que llegan a cruzar mil millones de eventos de uso procedentes de 6 editores, 4 agregadores y 4 grandes consorcios de bibliotecas que son recogidos en una inmensa base de datos. Este proyecto, en marcha desde 2006, también recomienda el uso de *OpenURL context object* como estándar para capturar y expresar ciertas partes de datos que son después transmitidos mediante *OAI-PMH* (**Bollen**; **Van de Sompel**; **Rodríguez**, 2008).

http://article-level-metrics.plos.org/

http://wiki.surffoundation.nl/display/standards/SURFshare+use+of+Usage+Statistics+Exchange

http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php

El *Open Access Statistik* del *DINI* en Alemania ha analizado, a nivel nacional, cómo normalizar la recogida de datos estadísticos poniendo el énfasis en la información procedente de los repositorios institucionales. Otras iniciativas interesadas en el desarrollo y promoción de estándares para el intercambio y generación de estadísticas agregadas son:

– el grupo de trabajo de estadísticas del *Knowledge Exchange Institutional Repositories* (*Knowledge Exchange*, 2007), una plataforma compuesta por el *JISC* británico, el *SURF* holandés, el *DFG* alemán y el *DeFF* danés; y

– el proyecto *PEER* (*Publishing and the ecology of European research*), financiado por la *Comisión Europea.*

http://www.dini.de/projekte/oa-statistik/english/

http://www.peerproject.eu/

2.2. Ventajas de disponer de estadísticas

Entre los motivos principales por los que en la última década hemos asistido a una explosión de repositorios destacan precisamente los beneficios que reportan en materia de visibilidad, impacto, uso y difusión de la producción científica.

La recolección sistemática de estadísticas puede ser una herramienta útil para que los repositorios puedan alcanzar objetivos internos y externos. El análisis del uso interno permite realizar un seguimiento de la producción científica depositada y estudiar así las pautas de crecimiento, ayudando a diseñar el plan de trabajo y las estrategias futuras para alimentar con más contenido. Reflejan también la visibilidad y la difusión internacionales y las tendencias de uso de estos archivos abiertos, que son indicadores de su eventual consolidación. Por otra parte las estadísticas pueden ser un medio persuasivo y elocuente para explicar el porqué de los repositorios abiertos ante la institución de la que dependen y su agencia financiadora –mostrando la relación coste-beneficio del repositorio– y ante la comunidad científica cuya investigación difunden y preservan –demostrando su efectividad en potenciar la accesibilidad de los resultados de investigación de un modo gratuito e inmediato en internet–.

Finalmente, los investigadores están interesados en acceder a estos datos para saber cuánta atención está recibiendo su investigación y cómo los usuarios están accediendo a este material, comparando el grado de "popularidad" de sus trabajos con el de otros compañeros de profesión (**Carr**; **Brody**; **Swan**, 2008).

"Los análisis estadísticos son un valor añadido para los administradores de los repositorios y para sus usuarios"

En resumen, los análisis estadísticos son un valor añadido para los administradores de los repositorios y para sus usuarios: miden su popularidad y uso, y contribuyen a la correcta toma de decisiones, a fijar las prioridades y a elaborar políticas mejores científicas.

Externamente a la institución pueden ayudar al desarrollo de nuevos marcos de evaluación científica de la producción de los investigadores, tal y como demuestra el *UK Research Excellence Framework* (*Research Excellence Framework*, 2009).

Recolectar los datos no es tarea fácil y una crítica recurrente es la

inexactitud y la ambigüedad que a veces conlleva su interpretación. Uno de los recelos predominantes se produce cuando no se sabe si se han filtrado o no la actividad de las arañas de los motores de búsqueda o las descargas de los robots, pues ello infla los resultados. Esta intromisión por parte de las máquinas es negativa, pues el objetivo es reflejar y examinar el verdadero uso académico recibido por el material científico presente en un repositorio. El filtrado del número total de clics, que incluye los dobles y múltiples clics y los generados por error o por azar, puede conseguirse mediante la identificación de usuarios y sesiones. Para hacer el seguimiento de un clic es necesario identificar previamente el usuario individual, lo que puede hacerse siguiendo la pista de su dirección IP. Sin embargo, esta opción plantea no sólo cuestiones de privacidad sino también de carácter técnico en los casos de los servidores proxy, IPs no resueltas o cuando hay una red de ordenadores que comparten una misma dirección IP. Alternativas pueden ser la activación de un sistema de identificadores por sesiones para realizar el seguimiento de cada visita o la activación de cookies.

Mientras se llega a un consenso en torno a esta cuestión fundamental, los repositorios de uno u otro software han experimentado recetas caseras (*DSpace ANU*[4], *Eprints Tasmania*[5]) o incorporando paquetes genéricos muy conocidos en la web. Uno de éstos es *Awstats*, difundido por *IRS Interoperable Repository Statistics*, un proyecto pionero de 2005 que pretendía crear un prototipo para capturar, usar e intercambiar datos estadísticos entre repositorios de *ePrints* y *DSpace* en modo consensuado. Entre los repositorios y otras plataformas abiertas que han propuesto aplicaciones en esta dirección cabe destacar las estadísticas de acceso del repositorio temático *RePEc LogEc* y las de acceso y citaciones de *Scielo*. *PLoS* han creado un sistema métrico de uso, citaciones y otros factores de impacto de los artículos de sus revistas, usando el prototipo de *Pirus*.

http://wiki.eprints.org/w/IRStats

http://logec.repec.org/

http://www.scielo.org/php/index.php?lang=es

Quedan algunos problemas técnicos y operativos debidos a la falta de estándares que impiden una completa interoperabilidad entre estadísticas generadas por repositorios a título individual (**Merk**; **Scholze**; **Windisch**, 2009):

– No existe acuerdo sobre cómo intercambiar esas estadísticas: como norma se generan gráficamente, no son exportables y en ocasiones su acceso es restringido a una comunidad preestablecida de usuarios.

– No hay acuerdo sobre qué se entiende por "uso", "event", "sesión", qué "set" de datos comparar (descargas de textos completos, visualización de metadatos, qué hacer con los registros con múltiples ficheros, etc.), cómo hacer tal comparación y cómo presentar los resultados. Las estadísticas de los repositorios pueden tener una amplia variedad tipológica (de uso, de depósitos, de acceso, de citaciones...) y pueden presentar grados de granularidad diversos. El acceso a los items además puede producirse por web server logs, por resolvedores y/o agregadores. La exclusión –o al menos, la señalización– de los accesos no humanos supone un consenso previo antes de la contabilización de datos.

– Las estadísticas contienen información de carácter personal y por ello se deben establecer políticas de privacidad así como políticas de estadísticas. Éstas, deben expresar su conformidad con las leyes locales de privacidad y protección de datos, esclarecer los usos y re-usos permitidos, definir quién controla la gestión de datos, diferenciar entre datos en acceso abierto y restringido, etc.

En España, el grupo de trabajo de estadísticas de *Recolecta*, en el que participa *Digital.CSIC*, se creó en 2009 para acercar los repositorios españoles a los esfuerzos de interoperabilidad en marcha en el panorama internacional. La agenda de trabajo del grupo para el año 2010 contempla un proyecto piloto que propondrá la normalización en la recolección de datos estadísticos (eliminando los dobles clics, la acción de robots, etc.) con el objetivo de comparar las estadísticas de uso de los repositorios españoles.

http://www.recolecta.net/wiki/index.php?title=Grupo_de_Trabajo_de_Estad%C3%ADsticas

3. Implementación y arquitectura de las estadísticas en *Digital.CSIC*

3.1. La motivación

La distribución territorial de los institutos del *CSIC* y sus bibliotecas ha motivado en gran medida la elaboración de las nuevas estadísticas de *Digital.CSIC*. Por el simple hecho de organizar, difundir y preservar de modo centralizado los resultados científicos de todos los centros e institutos de investigación del *CSIC* –centros que disfrutan de una considerable autonomía en planificación y gestión– *Digital.CSIC* es un repositorio con exigencias adicionales a las de otros cuya institución madre está dotada de una estructura más uniforme.

A pesar de este carácter centralizado, *Digital.CSIC* presenta una fuerte impronta distributiva: tras una fase inicial en la que su Oficina Técnica asumió íntegramente los primeros depósitos, en relativamente poco tiempo las bibliotecas se incorporaron a la tarea, subiendo las publicaciones de los investigadores de sus respectivos centros mediante el *Servicio de archivo delegado*

(*SAD*). El tercer pilar de este modelo son los propios investigadores, autorizados a depositar sus archivos personalmente. Con el paso de los meses se ha observado una tendencia hacia el aumento en la carga de depósitos por bibliotecas e investigadores, y una disminución relativa desde la Oficina Técnica, que puede así concentrar esfuerzos en dirigir cargas masivas automatizadas, explorar otras vías para garantizar la retroalimentación continua del repositorio y plantear otras líneas de acción.

Por ello, además de ayudar a la Oficina Técnica a realizar un seguimiento general del repositorio, se ha hecho evidente que las estadísticas necesitan llegar a un nivel de detalle importante no sólo en lo que respecta a las 9 áreas de investigación ("comunidades") presentes en *Digital.CSIC* sino también en lo que se refiere a cada uno de los centros e institutos del *CSIC* ("subcomunidades" en *DSpace*).

"Con las estadísticas las bibliotecas del CSIC realizan estudios y actividades de promoción y divulgación en sus centros"

En la segunda mitad de 2009 *Digital.CSIC* se planteó la mejora de su servicio de estadísticas, como respuesta a las peticiones de algunas bibliotecas de un nuevo módulo de datos por centros. En el pasado la Oficina Técnica había recibido puntualmente peticiones de informes estadísticos de algunos centros; estos análisis se realizaban manualmente, algo que resulta poco operativo de hacer con los más de 100 centros. Estas estadísticas a la carta nacen, pues, para que las bibliotecas del *CSIC* puedan llevar a cabo diversos estudios y actividades de promoción y divulgación en sus centros.

3.2. Aspectos técnicos del módulo de estadísticas

Arquitectura de *Digital.CSIC*

El repositorio se montó en 2007 sobre *DSpace*, un software desarrollado en sus comienzos por *HP* y el *MIT* sobre java (servlets, jsp, librerías de etiquetas jstl, etc.). Liberado en 2002 bajo una licencia *BSD* (*Berkeley software distribution*, para software de código abierto), es compatible con el protocolo *OAI-PMH*. A partir de 2009 ha pasado a formar parte del proyecto *DuraSpace.*

http://www.dspace.org

El primer diseño de *Digital.CSIC* comenzó a principios de 2008 sobre la versión 1.3.X de *DSpace* y a finales de ese mismo año salió a producción sobre la versión 1.4.X., que es la que está siendo usada actualmente.

DSpace funciona con las bases de datos *PostgreSQL* y *Oracle*. Se seleccionó *Oracle* como gestor de bases de datos porque es la recomendada por el *CSIC* y de la cual se pueden solicitar instancias de base de datos en alta disponibilidad para ser usadas rápidamente.

http://www.postgresql.org
http://www.oracle.com

El contenedor de servlets (servidor web para *JSP/Servlets*) utilizados para desplegar *Digital.CSIC* ha sido el software libre *Tomcat 6.X* recomendado por *DSpace* para la instalación de su aplicación. Para obtener un mejor rendimiento y facilitar la configuración se ha montado el servidor web *Apache 2.2.X* para servir htmls, documentos, imágenes, etc., y *Tomcat* únicamente para procesar *JSP/Servlets* etc. Ambos se comunican mediante el módulo *mod_proxy_ajp*.

http://httpd.apache.org
http://httpd.apache.org/docs/2.2/mod/mod_proxy_ajp.html

Evolución y elaboración de estadísticas en *Digital.CSIC*

El primer paso para obtener informes estadísticos sobre el uso y los depósitos en *Digital.CSIC* fue utilizar las estadísticas que genera *DSpace* si se configura para ello. *DSpace* incorpora una serie de servlets que son invocados cada vez que hay una petición al repositorio y que generan unos logs propios que luego pueden ser analizados, si fuera necesario, para preparar unos informes en html que pueden ser consultados por los usuarios con el rol "administrador". En general, estos informes estadísticos de *DSpace* son "pobres" en la información que ofrecen y no cubren las necesidades de *Digital.CSIC*.

El segundo paso fue intentar utilizar *Google Analytics* para obtener mejores estadísticas, pero surgieron problemas al intentar medir las descargas en el repositorio. *Google Analytics* utiliza un script que debe ser insertado en todas las páginas, ejecutándose al cargarse la página y registrando la visita, pero al hacer clic sobre un enlace que accede al fichero a texto completo no se carga ninguna página html/sjp y por tanto no se ejecuta el script y la descarga no queda registrada. Una posible solución es usar la función *pageTracker._trackPageview* del api de *Google Analytics* insertándolo en los enlaces que apuntan a los ficheros a descargar y así se consigue que la descarga quede registrada. Pero no es posible que los accesos directos a los ficheros (que no provienen de páginas del repositorio, sino de buscadores) queden contabilizados. Al no poder medir correctamente todas las descargas se terminó desechando también esta opción.

http://www.google.com/intl/es/analytics/

http://code.google.com/intl/es-ES/apis/analytics/docs/gaJS/gaJSApi.html

El tercer paso fue buscar módulos (add-ons) o aplicaciones de terceros que pudieran ser instalados en *Digital.CSIC* para obtener estadísticas más "ricas". El módulo *ePrints-Stats* de la *University of Tasmania*, cuya primera versión salió en mayo de 2007, era el que más se ajustaba a las necesidades, pues permitía un cierto grado de granularidad en la visualización de las estadísticas en función de algunos roles de *DSpace* y los informes (visitas a las páginas de metadatos y descargas de textos completos) cubrían las exigencias de *Digital.CSIC* en aquel momento.

http://www.dspace.org

http://www12.ocn.ne.jp/~zuki/Japanization/others/es-stats.html

Este módulo está desarrollado en java (servlets, jsp, jstl) y aprovecha las clases de *DSpace* y consultas *SQL(DML)* para realizar la conexión con la base de datos. Durante su instalación se vio un problema: el módulo sólo funciona con bases de datos *PostgreSQL*. Al utilizar *Oracle*, hubo que reescribir casi todas las sentencias *SQL* para hacerlo funcionar correctamente. Además se añadió la fuente de datos *GeoIP* (*GeoLite Country* de *MaxMind*) que permite saber las procedencia de cada visita en función de su IP.

http://www.maxmind.com/app/ip-location

El módulo de la *University of Tasmania* puede dividirse en dos partes: una que corre como una tarea programada y se encarga de analizar los logs y almacenarlos en tablas de la base de datos que, posteriormente, serán consultadas por la otra parte que lanza las consultas necesarias para poder visualizar los datos vía web. Con el módulo instalado y funcionando correctamente ya se disponía de estadísticas de visualizaciones y descargas generales y por registro (item) con la posibilidad de filtrar por roles de la aplicación o dejarlas en abierto y con la opción de consultarlas por meses y por años.

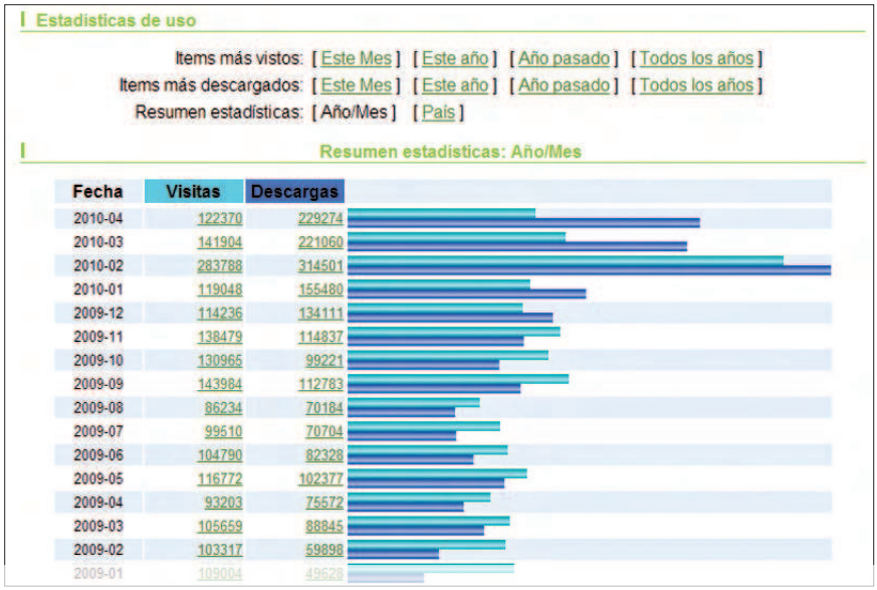

Gráfico 1. Estadísticas ofrecidas por el módulo ePrintsStats de la University of Tasmania

Con el último y cuarto paso se ha querido cubrir la demanda procedente de las bibliotecas y de la propia Oficina Técnica de *Digital.CSIC* de obtener informes aún más detallados e intentar ofrecer una mayor información estadística a los usuarios. Estas exigencias abrieron el camino para la elaboración de unas estadísticas a la carta a medida de las partes implicadas en el proyecto: centros e institutos de investigación, bibliotecas, Oficina Técnica e investigadores.

Con este nuevo módulo se siguen mostrando las estadísticas que se ofrecían desde el 2008 (con el módulo *ePrintsStats* de la *University of Tasmania*) a las que se han añadido las que se describen a continuación:

a) *Estadísticas generales:* Vienen a automatizar la generación y el enriquecimiento de informes que hasta la fecha la Oficina Técnica realizaba de manera manual y puntual bajo petición de las bibliotecas. Entre las novedades destacan la posibilidad de obtener un análisis sobre los tipos de registros y la opción de cruzar datos con la distribución territorial de los institutos del *CSIC*. Se pueden realizar búsquedas por período completo, año o mes concreto:

– número de centros/institutos por comunidades

– número de registros por comunidades

– número de registros por centros/institutos

– número de centros/institutos por distribución territorial

– número de registros por distribución territorial

– usuarios que depositan más registros

– autores con mayor número de registros

– tipos de registros

b) *Estadísticas por centros:* Surgen de la necesidad concreta de las bibliotecas del *CSIC* de informes exhaustivos. Suponen el máximo detalle estadístico que se ofrece y permiten búsquedas por un centro/instituto en período completo, año o mes concreto:

– número de registros depositados

– número de visualizaciones de registros

– número de descargas de registros

– usuarios por centro/instituto que más registros depositan

– títulos de los registros más visualizados (top20)

– títulos de los registros más descargados (top20)

– tipos de registros por centro

Número de Centros/Institutos por comunidades de Digital.CSIC

Comunidades de Digital.CSIC	Nº Centros/Institutos
Biología y Biomedicina	23
Ciencia y Tecnología de Alimentos	6
Ciencia y Tecnología de Materiales	10
Ciencia y Tecnologías Físicas	28
Ciencia y Tecnologías Químicas	14
Ciencias Agrarias	13
Humanidades y Ciencias Sociales	19
Recursos Naturales	21
Servicios Centrales CSIC	6

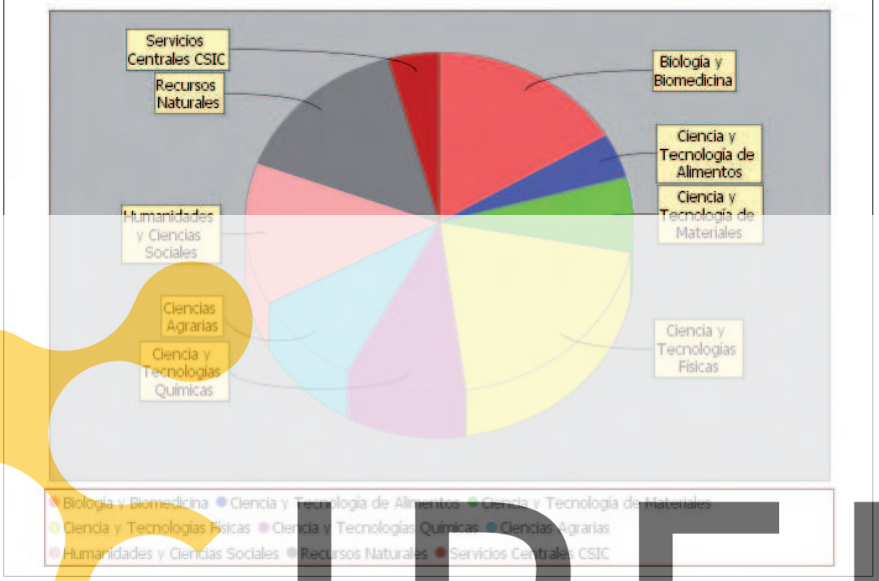

Gráfico 2. Número de centros/institutos por comunidades Digital.CSIC, abril de 2010

Para nutrir este módulo con los datos necesarios para ofrecer las nuevas estadísticas, se partía de las siguientes fuentes de información:

– Datos obtenidos del módulo de la *University of Tasmania* sobre el análisis de los logs de accesos de *Digital.CSIC* y que almacena posteriormente en la base de datos.

– Información que *DSpace* almacena en su base de datos sobre comunidades, colecciones, depósitos, etc. (para su funcionamiento interno).

– Tablas maestras auxiliares creadas para mostrar la relación de centros, comunidades con su provincia, etc.

El módulo muestra las estadísticas arriba indicadas en el período seleccionado por el usuario realizando las consultas adecuadas sobre estas 3 fuentes que están repartidas en distintas tablas en la misma base de datos.

Para seguir con la misma línea de programación que la aplicación sobre la que se ha montado (*DSpace*), el nuevo módulo de estadísticas se ha implementado en java (servlets, jsp, librerías de etiquetas jstl), utilizando las clases ya existentes en *DSpace* dentro del paquete *org.dspace.storage.rdbms* para operar con la base de datos mediante un pool de conexiones, generar consultas, obtener *result sets,* etc. Además, se han creado filtros por roles de la aplicación (algunos de ellos nuevos) para que sólo determinadas personas (administradores del repositorio, el personal de bibliotecas del centro en cuestión) pudieran acceder a datos de carácter personal como "Usuarios que depositan más registros". La creación de estos filtros en las nuevas estadísticas va en sintonía con la política de estadísticas y la de privacidad de *Digital.CSIC*.

http://java.sun.com/products/jsp/jstl/

Por último, para generar gráficas "vistosas" vía web en el mismo momento de la petición y en función de las consultas realizadas, se utiliza la librería *JFreeChart* (un framework de código abierto para java que permite generar distintos tipos de gráficas).

http://www.jfree.org/jfreechart/

Conclusiones 1.0

En mayo de 2010 se hizo pública la versión 1.0 del nuevo módulo de estadísticas del repositorio *Digital.CSIC*. Se pretende continuar con él, integrando datos de proyectos que están en marcha y que permitirán realizar seguimientos exhaustivos de la producción científica del *CSIC*. Por otro lado, la Oficina Técnica recoge las sugerencias de las bibliotecas e investigadores del *CSIC* para dotar el módulo de mayor funcionalidad y crear informes más completos en versiones futuras.

La posibilidad de realizar búsquedas avanzadas sobre el módulo, la vinculación de datos estadísticos con los factores de impacto de los trabajos del *CSIC* presentes en el repositorio, una mayor granularidad en los datos, la integración con información sobre producción científica procedente de otras fuentes (otras bases de datos del *CSIC*, otros repositorios, plataformas editoriales etc), y la adhesión a estándares y protocolos internacionales son medidas necesarias en aras de una verdadera interoperabilidad entre datos que reflejen el uso y el impacto de la comunicación científica en un entorno de acceso abierto.

A nivel de estándares internacionales, cabe esperar ver los resultados del proyecto *Pirus2* a final de este año 2010, ya que ha despertado expectativas sobre la posibilidad de crear un prototipo que permita crear estadísticas de uso agregadas sobre artículos individuales que pueda ser usado tanto por repositorios (institucionales y temáticos) como por editores/agregadores. Este prototipo en xml podría suponer el desarrollo de *Counter* como estándar de

Registros por Centro/Instituto más descargados (TOP20)

Títulos	Descargas
- SPEIbase: a global 0.5º gridded SPEI data base (plain text)	11445
- SPEIbase: a global 0.5º gridded SPEI data base (raw binary)	6200
- Frutas y verduras congeladas	4198
- Genética aplicada	2316
- Respuesta fisiológica, bioquímica y molecular de las leguminosas a estreses abióticos	2292
- SPEIbase: a global 0.5º gridded SPEI data base (I. NetCDF)	2286
- Caracterización fisiológica del Bitter Pit: aspectos nutricionales, fenológicos y de diagnóstico	2012
- Veza común, su cultivo y utilización	2008
- Desarrollo de un sistema de recolección mecánica de aceituna en contínuo.	1857
- El ciclo de las xantofilas en plantas sometidas a estrés nutricional o climático	1607
- Uniformidad de los sistemas de riego por aspersión en el polígono de la Loma de Quinto (Zaragoza)	1537
- CIERZO - Variedad de cebada	1526
- Genética : su posición entre las ciencias biológicas	1417
- Les groupements vegetaux du bassin moyen de l'Ebre et leur dynamisme	1353
- Incompatibilidad entre patrón e injerto. Variedades de ciruelo injertadas sobre híbridos almendro x melocotonero	1331
- El calcio nutriente para las plantas. Bitter pit en manzano	1313
- Plantación de frutales	1282
- SPEI Calculator	1256
- Evaluación del riego por aspersión en cobertura total en una parcela de maiz	1225
- Revisión de las necesidades hídricas netas de los cultivos de la Cuenca del Ebro	1222

Gráfico 3. Registros más descargados para la Estación Experimental de Aula Dei (EEAD), abril de 2010

facto y la promoción de *OpenURL context object* como el estándar por el cual los repositorios expresen sus datos estadísticos de uso tanto a plataformas externas como a los servidores locales. Los datos estadísticos de los repositorios pueden ser recolectados mediante el protocolo *OAI-PMH* y como usan el formato xml pueden ser capturados por *Sushi*, el protocolo usado por los editores para intercambiar estadísticas. Otra ventaja de *OpenURL context object* es su capacidad para ser extensible y comprimir datos.

Para poder realizar estadísticas de uso agregadas se plantea un problema: los artículos son accesibles desde una variedad de plataformas (de editores, de agregadores, de repositorios, mediante resolvedores, etc.) y por tanto para la generación de datos estadísticos se necesita un consenso sobre el identificador de los mismos. En la práctica, el identificador más común es el *DOI*, pero en los repositorios hay una gran variedad de campos de metadatos en que este identificador puede aparecer (en los campos de descripción, de relaciones, de derechos, de identificadores, etc). Por otra parte, es necesario llegar a un consenso sobre otras cuestiones para poder comparar estadísticas de diferentes plataformas y evitar duplicados: por ejemplo, el procedimiento para identificar –y, por tanto, medir– a los usuarios de repositorios (lo que puede plantear problemas con las políticas de privacidad); el tipo de sesión objeto del análisis (full-text, página principal, peticiones con éxito y con error...); el lapso temporal asociado a las sesiones de uso; el tipo de acceso (humano o generado automáticamente), etc.

Aparte de las perspectivas concretas que se abren para el análisis agregado del uso e impacto de los artículos individuales, los repositorios albergan una variedad más amplia y rica de items, que deben ser incorporados en sus estadísticas locales. Es éste un servicio de valor añadido de los repositorios, ya que engloban contenido científico que cae fuera del campo de acción de los editores. Se pretende mejorar las estadísticas por centros y las generales para poder analizar las dinámicas de uso de los registros por tipos documentales y áreas científicas y ofrecer listados por autores más descargados y visitados y no sólo por títulos de documentos; la exportación de resultados en diferentes formatos es otro servicio para poner en práctica y la integración de diversas fuentes de información del *CSIC* sobre la producción científica de su comunidad de investigadores permitirá la incorporación de criterios métricos diferentes, como son los de impacto y citaciones. Por supuesto, el desarrollo de unas estadísticas complejas que hagan uso de diferentes tipos de items así como de criterios de medición interrelacionados se sustenta en unas políticas de estadísticas y de privacidad bien definidas.

Agradecimientos

A Luisa Domènech (*Digital. CSIC*), Carmela Pérez-Montes (*Biblioteca Tomás Navarro Tomás, Centro de Ciencias Humanas y Sociales, CSIC*), José-Carlos Martí-

nez-Giménez (*Biblioteca de la Estación Experimental de Aula Dei, CSIC*) y Pablo De-Castro (Servicio de Recursos Electrónicos de la Biblioteca, *Universidad Carlos III de Madrid*) por sus contribuciones en la implementación del nuevo modulo de estadísticas de *Digital.CSIC*.

Notas

1. *Consejo Superior de Investigaciones Científicas* (*CSIC*). *Plan de actuación del CSIC 2010-2013*, octubre de 2009.
http://www.csic.es/web/guest/plan-de-actuacion-2010-2013

2. *Consejo Superior de Investigaciones Científicas* (*CSIC*). *Memoria del CSIC 2009*, junio de 2010.
http://www.csic.es/web/guest/memorias

3. *Oficina Técnica de Digital.CSIC. Memoria 2009*, abril de 2010.
http://digital.csic.es/handle/10261/23383

4. Estadísticas de *DSpace ANU*
http://sts.anu.edu.au/drs/downloads/dspace-stats/readme.html

5. Estadísticas elaboradas por la *University of Tasmania* para *ePrints*
http://eprints.utas.edu.au/262/

Bibliografía

Aguillo, Isidro. "Acceso abierto. Una nueva generación de métricas e indicadores". En: *Fesabid 2009, XI Jornadas españolas de documentación*, 20-22 de mayo 2009.
http://www.slideshare.net/fesabid/aguillo-acceso-abierto-una-nueva-generacion-de-metricas-e-indicadores

Apache
http://httpd.apache.org

Bollen, Johan; **Rodriguez, Marko A.**; **Van-De-Sompel, Herbert**;. "Towards usage-based impact metrics: first results from the *Mesur* project". En: *Proc. of the 8th ACM/IEEE-CS joint conference on digital libraries*, Pittsburgh, 16 de junio 2008.
http://arxiv.org/abs/0804.3791

Bollen, Johan; **Van-De-Sompel, Herbert**; **Rodriguez, Marko A.** *Mesur*: usage-based metrics of scholarly impact. In: *JCDL '07* Vancouver, CA.
http://www.mesur.org/Documentation_files/JCDL07_bollen.pdf

Carr, Leslie; **Brody, Tim**; **Swan, Alma**. "Repository statistics: what do we want to know?" En: *Third intl conf on open repositories*, 2008, Southampton, 1-4 abril de 2008.
http://pubs.or08.ecs.soton.ac.uk/30/

Consejo Superior de Investigaciones Científicas (*CSIC*). *Plan de Actuación del CSIC 2010-2013*, octubre de 2009.
http://www.csic.es/web/guest/plan-de-actuacion-2010-2013

Consejo Superior de Investigaciones Científicas (*CSIC*). *Memoria CSIC de 2009*, junio de 2010.
http://www.csic.es/web/guest/memorias

Counter project
http://www.projectcounter.org/

DINI Open Access Statistik
http://www.dini.de/projekte/oa-statistik/english/

DOI
http://www.doi.org/

DSpace
http://www.dspace.org

ePrintsStats, University of Tasmania
http://www12.ocn.ne.jp/~zuki/Japanization/others/es-stats.html

GeoIP
http://www.maxmind.com/app/ip-location

Google Analytics
http://www.google.com/intl/es/analytics/

Google Analytics Tracking API
http://code.google.com/intl/es-ES/apis/analytics/docs/gaJS/gaJSApi.html

Grupo de estadísticas de *Recolecta*
http://www.recolecta.net/wiki/index.php?title=Grupo_de_Trabajo_de_Estad%C3%ADsticas

Herb, Ulrich; **Kranz, Eva**; **Leidinger, Tobias**; **Mittlesdorf, Bjorn**. "How to assess the impact of an electronic document? And what does impact mean anyway? Reliable usage statistics in heterogeneous repository communities". *OCLC systems & services*, 2010, v. 26, n. 2, pp. 133-145.
http://scidok.sulb.uni-saarland.de/volltexte/2010/3158/

IRStats software
http://wiki.eprints.org/w/IRStats

JFreeChart
http://www.jfree.org/jfreechart/

Knowledge Exchange Usage statistics workshop Strand report. Strand title: Usage statistics, febrero de 2007.
http://www.driver-repository.be/media/docs/KEIRstrandreportUsageStatisticsFINALFeb07.pdf

Librerías de etiquetas *JSTL*
http://java.sun.com/products/jsp/jstl/

Merk, Christine; **Scholze, Frank**; **Windisch, Nils**. "Item-level usage statistics: a review of current practices and recommendations for normalization and exchange". *Library high tech journal*, 2009, v. 27, n. 1, pp. 151–162.
http://elib.uni-stuttgart.de/opus/volltexte/2009/4115/index.html

Módulo mod_proxy_ajp
http://httpd.apache.org/docs/2.2/mod/mod_proxy_ajp.html

Módulos para *Dspace* (add-ons)
http://www.dspace.org/add-ons-and-extensions/addons/

OAI-ORE
http://www.openarchives.org/ore

OAI-PMH
http://www.openarchives.org

Oficina Técnica de Digital.CSIC. Memoria 2009, abril de 2010.
http://digital.csic.es/handle/10261/23383

OpenURL context object
http://www.niso.org/kst/reports/standards?step=2&project_key=d5320409c5160be4697dc046613f71b9a773cd9e

Oracle
http://www.oracle.com

PEER
http://www.peerproject.eu/

PLoS article level metrics
http://article-level-metrics.plos.org/

PostgreSQL
http://www.postgresql.org

Publish or perish
http://www.harzing.com/pop.htm

Ranking Webometrics 2010 de centros de investigación, enero de 2010.
http://research.webometrics.info/top4000_r&d_es.asp

RePEc LogEc
http://logec.repec.org/

Research Excellence Framework. A brief guide to the proposals, octubre de 2009
http://www.hefce.ac.uk/research/ref/resources/REFguide.pdf

Scielo
http://www.scielo.org/php/index.php?lang=es

Scimago Institutions Rankings 2009
http://www.scimagoir.com/

Scholze, Frank. "Measuring research impact in an open access environment". *Liber quaterly*, 2007, v. 17, n. 3/4.
http://elib.uni-stuttgart.de/opus/volltexte/2007/3234/

Shepherd, Peter. "The Counter code of practice and ICOLC Europe". Paris, 27 octubre de 2009.
http://www.projectcounter.org/documents/ICOLCEuropeOct2009.ppt

Shepherd, Peter; **Needham, Paul**. *Pirus final report*, enero de 2009.
http://www.jisc.ac.uk/media/documents/programmes/pals3/pirus_finalreport.pdf

SurfShare SURE
http://wiki.surffoundation.nl/display/standards/SURFshare+use+of+Usage+Statistics+Exchange

Sushi
http://www.niso.org/workrooms/sushi

The Pirus2 Project. Project plan and progress.
http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php

***Isabel Bernal-Martínez, Julio Pemau-Alonso**. Consejo Superior de Investigaciones Científicas, Unidad de Coordinación de Bibliotecas.*
isabel.bernal@bib.csic.es
julio.pemau@bib.csic.es

¡¡ Nada más...

ACTUALIDAD BIBLIOTECARIA
BIBLIOTECAS PÚBLICAS
NUEVAS TECNOLOGÍAS
BIBLIOTECAS UNIVERSITARIAS
GESTIÓN DOCUMENTAL
OTROS ESPACIOS DE LECTURA
CON PASADO Y CON FUTURO
BIBLIOTECARIOS INSIGNES
AGENDA DE FORMACIÓN
BIBLIOTECAS ESCOLARES
LAS BIBLIOTECAS DE MI VIDA
RECOMENDACIONES DE LECTURA

...y nada menos !!

Mi Biblioteca

La revista del mundo bibliotecario

Suscríbete a *Mi Biblioteca* y recibirás cada año, de manera gratuita, el *Calendario de la Lectura* y el *Anuario de Bibliotecas Españolas* de la Fundación Alonso Quijano.

Tfno. 952 23 54 05
www.mibiblioteca.org